GeneSV - approach to help characterize possible variations in genomic and protein sequences

A. Zemla, T. Vasilevska, E. Volkova, D. W. C. Beasley, S. C. Weaver, N. Vasilakis, P. Naraghi-Arani

August 22, 2012

Bioinformatics and Biology Insights

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

GeneSV – an approach to help characterize possible variations in genomic and protein sequences

Adam Zemla[1*], Tanya Kostova[1], Rodion Gorchakov[2,3,4], Evgeniya Volkova[2,3,4], David W. C. Beasley[2,3,4,5], Jane Cardosa[6], Scott C. Weaver[2,3,4], Nikos Vasilakis[2,3,4], Pejman Naraghi-Arani[7*]

[1] Computing Application & Research Department, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA
[2] Department of Pathology and Center for Biodefense and Emerging Infectious Diseases, University of Texas Medical Branch, Galveston, TX 77555-0609, USA
[3] Center for Tropical Diseases, University of Texas Medical Branch, Galveston, TX 77555-0609
[4] Institute for Human Infections and Immunity, University of Texas Medical Branch, Galveston, TX 77555-0610, USA
[5] Department of Microbiology and Immunology, University of Texas Medical Branch, Galveston, TX 77555, USA
[6] Sentinext Therapeutics Sdn Bhd, 10050 Penang, Malaysia
[7] Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA

*To whom correspondence should be addressed: Pejman Naraghi-Arani (Tel: 1-925-422-5665; FAX: 1-925-422-2282; Email: naraghiarani2@llnl.gov) or Adam Zemla (Tel: 1-925-423-5571; FAX: 1-925- 423-6437; Email: zemla1@llnl.gov)

ABSTRACT

A computational approach for identification and assessment of genomic sequence variability (GeneSV) is described. For a given nucleotide sequence, GeneSV collects information about the permissible nucleotide variability (changes that potentially preserve function) observed in corresponding regions in genomic sequences, and combines it with conservation/variability results from protein sequence and structure-based analyses of evaluated protein coding regions. Results generated by the system may: **(a)** contribute to functional annotation of genes or genomes; **(b)** predict potential mutations not observed in current databases; **(c)** help estimate mutation rates for genomes with high plasticity; **(d)** help assess the accuracy of novel mutation positions reported from sequencing efforts. GeneSV was used to predict effects (functional vs. non-functional) of 37 amino acid substitutions on the NS5 polymerase (RdRp) of dengue virus type 2 (DENV-2), 36 of which are not observed in any publicly available DENV-2 sequence. Thirty two novel mutants with single amino acid substitutions in the RdRp were generated using a DENV-2 reverse genetics system. In 81% (26 of 32) of predictions tested, GeneSV correctly predicted viability of introduced mutations. In 4 of 5 (80%) mutants with double amino acid substitutions proximal in structure to one another GeneSV was also correct in its predictions. Predictive capabilities of the developed system were illustrated on dengue RNA virus, but described in the manuscript a general approach to characterize real or theoretically possible variations in genomic and protein sequences can be applied to any organism.

INTRODUCTION

RNA viruses have exceptionally high mutation rates [1-2], which enable them to form mixed variant virus populations referred to by many authors (e.g. [3-5]) as “quasispecies.” The high genetic variability within the quasispecies facilitates virus adaptation to different environments and hosts and overall fitness [6-8]. Public databases such as GenBank [9], the European Nucleotide Archive (ENA) [10], the Universal Protein Resource (UniProt) [11], and the Protein Data Bank (PDB) [12], contain a growing volume of sequence and structural information, which encompasses only a part of the genetic diversity of viral species. Further, these databases are currently biased towards the consensus genotypes in clinical isolates from symptomatic cases. It is unlikely that the full array of viable viral genotypes of a given species will ever be represented in these databases due to sampling biases and the fluid nature of the viral quasispecies clouds that are present in each infection event. Yet, in certain cases (such as attempts to generate attenuated strains of viruses via molecular cloning for vaccine development), it has become important to evaluate the effects of sequence changes on the viability of the virus. While this can be done via reverse engineering and fitness experiments, an *in silico* predictive system that can produce such evaluations would be a valuable (and more rapid) guiding tool that should enhance the rates of success for such work.

A number of computational methods have been developed to predict the functional effect of a non-synonymous single-nucleotide polymorphism (nsSNP), a single-nucleotide change in a protein-coding region of a gene that causes an amino acid substitution (AAS) in the resulting protein. A review of existing computational approaches to estimate the deleteriousness of single nucleotide variants has been recently published by G.M. Cooper and J. Shendure [13]. The most common approaches to estimate deleteriousness exploit the fact that sequences observed among living organisms are those that have not been removed by natural selection. Hence, homology searches and conservation analysis are two main components of a majority of such predictive systems. Examples of widely used methods are SIFT (Sorting Tolerant From Intolerant) [14], PROVEAN (Protein Variation Effect Analyzer) [15], and PolyPhen-2 (Polymorphism Phenotyping v2) [16]. For a given query protein sequence, SIFT performs searches for functionally related protein sequences using the PSI-BLAST algorithm [17]. From calculated multiple sequence alignments it assesses probabilities of all theoretically possible 20 amino acids substitutions at each position in the query sequence. In the SIFT scoring system the positions identified as highly conserved are considered as intolerant to most substitutions, whereas poorly conserved positions are expected to tolerate substitutions. In comparison with the SIFT method, the PROVEAN algorithm predicts the functional impact for all classes of protein sequence variations, not only single amino acid substitutions but also insertions, deletions, and multiple substitutions. To collect homologous and related sequences from the NCBI NR protein database [9] it uses the BLASTP algorithm [17]. From calculated pairwise sequence alignments PROVEAN extracts alignment scores between an unmutated query protein sequence, a query sequence with introduced amino acid variants, and closely related homologous proteins that are known to be functional. The conservation scoring system relies on calculated “delta scores” – differences in estimated alignment scores between query and homologs, and query variants and homologs. Amino acid substitutions introduced into query variants that reduce the similarity (calculated using substitution matrix BLOSUM62) of protein sequence to the functional homologs are considered to cause a damaging effect. PolyPhen-2 differs from SIFT and PROVEAN as it was specifically designed to estimate the deleterious nature of human genetic variants. It uses machine learning methods to select the optimal set of features (sequence- and structure-based) to predict the possible impact of an amino acid substitution on the structure and function of human proteins. The choice of homologous protein sequences is done by BLAST searches from which multiple sequence alignments are calculated using the MAFFT program [18]. For the PolyPhen-2 development two pairs of datasets were used to train and test the scoring system. Each pair consisted of a set of proteins with known human disease-causing mutations, and a set of human or mammalian proteins without annotated involvement in disease. Direct comparison of PolyPhen-2 with PROVEAN on similar human and non-human protein variant datasets showed that both methods perform with similar accuracy of 78–79% [15].

Here, we describe a new method, the GeneSV approach, which helps characterize real or theoretically possible variations in genomic and protein sequences. To assess viability of possible mutations our system combines information from all three types of data: genomic sequences, protein sequences, and protein structures. Homology searches and final sequence alignments are calculated using PSI-BLAST and Smith-Waterman algorithms [19], respectively. Constructed structural models are used to refine identification of structural homologs and to distinguish structurally conserved from structurally variable regions in homologous proteins. The final scoring system, a measure of confidence that the point mutations under consideration will define a viable or non-viable virus, is a combination of scores that are based on both identified sequence and structure features. The performance of the GeneSV system was demonstrated by predicting the viability of dengue virus type 2 (DENV-2) mutants that, at the time of performing the predictions and validation experiments, had not yet been reported in GenBank or ENA. Using results from the biological experiments, we evaluated the accuracy of created predictions and compared the GeneSV results with predictions made by other methods: SIFT, and PROVEAN. A selection of mutation positions for this study (10 codons and 37 mutants) was made in order to test the accuracy of GeneSV and other prediction systems when the confidence level in generated predictions for such positions could be low. For this, we focused on positions and protein regions that were outside of known active sites and chose amino acid substitutions that have not been seen in available sequence databases. All selections and computational predictions were made prior to any experimental validation. Details of performed selections and predictions are described in the Material and Method section (see “Selection of position mutations for the GeneSV characterization and experimental validation.”)

MATERIALS AND METHODS

1. Description of the GeneSV approach

In its approach to help characterize possible variations in genomic and protein sequences within a species, the GeneSV system uses a known sequence space for a given genome (where a genome can be defined as broadly as a complete genome or as narrowly as a single gene) in combination with sequence and structural analyses performed at the protein level. The reported results can be used to infer the probability of existence of novel mutations (new sequence variants) that result in a viable protein or organism (e.g. virus.) GeneSV evaluates the probability of selected mutation(s) first by analyzing its (their) frequency (frequencies) observed in the genomic sequence databases. The evaluation is further refined using results from automated protein sequence homology searches, protein structure modeling and structure similarity analyses, an assessment of functional features assigned to the specific regions in genes or a genome, and the calculated theoretical probability of codon mutations.

a) *Analysis of observed nucleotide and codon frequencies and theoretical probability of codon mutation*. For a given base and codon in a genomic sequence (e.g. RefSeq – reference genomic sequence for a given organism) an analysis of the observed frequencies is performed using calculated alignments with all genomic sequences from available databases (e.g. ENA or GenBank). From this analysis, we construct the *codon frequency matrix* and the base *position frequency matrix,* which represent the numbers of occurrence of each of the 64 codons and of the 4 bases at each position in the genome sequence. In addition to the analysis of observed codons we also estimate the likelihood that a given codon can be constructed from the calculated base position frequency matrix. Further assessment of possible position changes includes the analysis of the theoretical probability of codon mutations, e.g. number of transitions or transversions (see subsections below).

b) *Sequence-based analysis of possible sequence variabilities at the protein level* is performed using our SeqalSV (sequence alignment-based sequence variability) module implemented within the AS2TS homology-based protein structure modeling system [20]. For a given protein coding region, SeqalSV performs protein homology searches against protein sequence databases (e.g. UniProt). The residue-residue correspondences derived from the calculated alignments allow construction of sequence motifs and profiles important for assessing regions of residue conservation. This information facilitates construction and refinement of structural homology models, and assists protein function predictions. Our approach to identify residue conservation among proteins differs from most of the currently used methods that rely on sequence-based comparisons only (e.g. calculated multiple sequence alignments using PSI-BLAST, ClustalW, Muscle, or other sequence alignment programs). A serious limitation of strictly sequence similarity-based approaches is that there is typically little or no confidence in assigning residue-residue correspondences among proteins when the level of sequence identity between the compared proteins is poor. If structural models are available, then structure comparison algorithms may provide much higher confidence in assigning residue-residue correspondences than sequence-based algorithms alone. Nevertheless, even calculated structural alignments, if standard structure alignment procedures are applied, may be inaccurate: for some compared proteins, or regions therein, more than one possible superposition can reasonably be reported, and it may be difficult to decide which alignment is most satisfactory. In our approach of calculating residue-residue correspondences we explore results from sequence, structure, and local and global alignment calculations as described below.

c) *Combined sequence and structure based analyses*. To enhance the confidence in the calculated residue-residue correspondences we combine sequence and structure-based analyses. We use the StralSV (structure alignment-based sequence variability; [21]) algorithm to perform comparisons between a reference protein structure and proteins in a structure database through detection of closely related structure fragments and quantification of residue frequency from tight local structure alignments. The SeqalSV sequence variability evaluation system complements results from StralSV by performing comparisons between a reference protein sequence and proteins in the sequence databases. Both algorithms are used together in a protocol in which a given protein is analyzed to determine which of its residue positions are broadly or uniquely conserved when compared to large, representative samplings of proteins. Structural models for StralSV calculations are taken (if available) from Protein Data Bank (PDB) or constructed using AS2TS homology-based protein structure modeling system. We also use structural models to identify highly divergent homology with proteins (those not detectable through sequence-based searches alone) and to detect protein fragments in other proteins from PDB that may share structural similarity only in local functional regions. For this task, we use a structure alignment algorithm LGA [22], which is a component of the StralSV system. The LGA program performs structure-based alignment calculations between the query proteins and the fragments found in structures from PDB, and quantifies the sequence and structure variabilities at each residue position. This approach as implemented within the StralSV system allows identification of invariant residues (often essential to protein function), unusual variants and regions, which tolerate sequence variation. The constructed structural models are also used in the characterization of observed mutation points by assigning their location on a given protein (e.g. buried, exposed, within active site, part of predicted antigenic determinants, being in contact with other residues, etc.), or assessing how they cluster together with corresponding residues in other homologous proteins from related organisms. Such findings allow making predictions about possible new mutations that are not yet observed for a given organism in current sequence and structure databases. This approach can help discover conserved fragments and even conserved single residues that are unique for a virus species, as well as strongly conserved fragments characterizing a family of functional proteins. It can be used to guide experiments and propose specific regions to apply selective pressure. This type of analysis also gives guidance in identifying compensatory mutations, which are of great importance to predicting evolutionary pathways.

2. Data collection and processing

We have selected a stable Full-Length Infectious Clone (FLIC) of DENV-2 (gi:132271146; derived from a sylvatic DENV-2 strain P8-1407 isolated in Malaysia in 1970; [23],[24]) as an input sequence to illustrate the capabilities of the GeneSV approach with regards to point mutation predictions and their characterization. The following steps in collecting and processing data have been performed:

1. We started data collection by selecting a complete genome of DENV-2 RefSeq (gi:158976983) as a reference sequence for GeneSV analysis. The RefSeq sequence was used to calculate position frequency matrices, detection of new mutation positions, and to estimate distances from RefSeq to the FLIC and RefSeq to new mutants we planned on introducing into the FLIC. In this study, we used RefSeq instead of the consensus sequence because the differences between RefSeq and the consensus were not significant.
2. A complete set of available genomic sequences for all DENV serotypes was downloaded (date: 2012.02.12) from two databases: GenBank and ENA.
3. The collected genomic sequences were processed to remove redundancy (i.e. duplicated sequences). A non-redundant set of all Dengue sequences (DENV-1234) from a constructed dataset 2012.02.12 consists of 11,964 sequences:
   - 3,809 sequences for DENV-1
   - 3,586 sequences for DENV-2
   - 3,132 sequences for DENV-3
   - 941 sequences for DENV-4
   - 496 Dengue sequences with no specified type
4. Position and codon frequency matrices using RefSeq and FLIC were calculated for both the DENV-2 sequence library, and also for the combined DENV-1234 sequence library of all collected sequences. Codons or base positions with fewer than 3 position hits were marked by (*) to help refine our level of confidence in the GeneSV assessments produced.
5. Sequence variability and conservation analysis (SeqalSV) for RefSeq and FLIC proteins against UniRef_100 library of protein sequences from UniProt was completed as an automated procedure within the AS2TS system. Sequence-structure variability and conservation analysis (StralSV) was performed on all structural models constructed. Results from protein sequence (SeqalSV) and structure (StralSV) variability calculations were completed and combined with position and codon frequency matrices of evaluated genomic sequences.
6. Structural models for RefSeq and FLIC proteins were constructed using the AS2TS system. The highest accuracy structural models were generated for NS5, NS3, and Envelope proteins.
7. Ten candidate positions in the RdRP nucleotide sequence from FLIC DENV-2 were selected, 37 codon substitutions proposed, and predictions of resulting changes in the virus growth (viability of

the mutant) made. The viability of the constructed mutants was later experimentally tested to evaluate computational predictions.

8. A second complete set of available genomic sequences for all Dengue serotypes was downloaded from GenBank and ENA on February 18th, 2013 (one year after the first set was downloaded) in order to assess the possible effects of dataset changes on created predictions. All collected genomic sequences were processed to remove redundancy. A non-redundant set of all Dengue sequences (DENV-1234) from the dataset 2013.02.18 consisted of 13,302 sequences:
   - 4,347 sequences for DENV-1
   - 3,985 sequences for DENV-2
   - 3,286 sequences for DENV-3
   - 1100 sequences for DENV-4
   - 584 Dengue sequences with no specified type

The two datasets (2012.02.12 and 2013.02.18) were compared, and 1205 new mutations were identified in DENV-2 genomic sequences. Of these, 1170 novel codon positions were found in the coding regions and 35 new base positions in the non-coding regions (5' and 3' non-translated regions). The identified 1205 new mutation positions and the older dataset of genomic sequences (2012.02.12) were submitted to the GeneSV system to test its efficacy in predicting novel viable mutations and to measure the robustness of the system by assessing the impact of these newly discovered mutations on the predictions already made.

3. Description of the assessment methods and scoring schemes implemented within the system

Assessment of the probability that a given mutation of one codon into another would produce a viable variant member of the quasispecies is based on the analysis of data from the genomic sequences, the protein sequence homology searches, and structural modeling and sequence-structure conservation analysis of corresponding protein regions. In the GeneSV system, each proposed codon mutation is initially characterized using nine criteria, which are based on observed similarities with corresponding positions in genomic and protein sequences from available databases. The input for this processing comprises two selected genomic sequence databases called reference library (e.g. library of DENV-2 genomic sequences) and expanded library (e.g. library of DENV-1, -2, -3, and -4 genomic sequences), and one protein sequence library (e.g. UniRef_100 from UniProt). To each mutation position of interest (e.g. codon) in a given reference sequence the following criteria are assigned:

O0) codon is present in the reference sequence library

O1) codon can be constructed from base position frequency matrix from the reference library

O2) codon represents amino acid (AA) which is observed at the corresponding position in sequence from the reference library

O3) codon can be constructed from additional bases inferred from observed AA at the corresponding position in sequence from the reference library

O4) codon is present in observed set of codons in expanded library

O5) codon can be constructed from observed bases from expanded library

O6) codon represents amino acid (AA) observed in expanded library

O7) codon represents amino acid (AA) observed in combined genomic and protein sequence libraries (expanded library of genomic sequences + library of homologous protein sequences)

O8) codon can be constructed from all additional bases inferred from amino acids (see O7) observed at corresponding AA positions from combined genomic and protein sequence libraries

In the GeneSV system, the observation-based characteristics listed above are reported for each evaluated codon in a format of the binary matrix as shown in Table 3 (examples of complete "summary" file outputs from the system are provided in the supplemental data – Suppl. 1, 2, and 3). The characteristics ("O0" to "O8") are used to define different "confidence scores" that can be assigned to predictions of the sequence changes that are likely allowable or likely disallowed from a functional standpoint. These confidence scores were used to predict the likely viability of various NS5 mutants that may be observed and from which we chose the mutations we introduced into the NS5 gene of the FLIC. The currently proposed seven levels of confidence scores (viable: V1-V4, non-viable: N5-N7) that are assigned to predictions and assessment of possible sequence changes are as follows:

(V1) VIABLE - Highest confidence for viable mutations is assigned based on codons observed in the reference library (e.g. DENV-2 codon frequency matrix). (Observation O0).

(V2) VIABLE - High confidence for viable mutations (novel codon predictions) is assigned based on codons observed in the expanded library (e.g. DENV-1234 codon frequency matrix). (Observation O4).

(V3) VIABLE - Medium confidence for viable mutations (novel codon predictions) is assigned to codons that can be constructed from the position frequency matrix from the expanded library and the AA coded for is observed at the corresponding position in synonymous codons or in results from protein sequence/structure homology searches (SeqalSV/StralSV analyses). (Observations O5 and O7).

(V4) VIABLE - Low confidence for viable mutations (novel codon predictions) is assigned to codons that cannot be constructed from the position frequency matrix from expanded library, but can be predicted based on corresponding AA positions from synonymous codons or SeqalSV/StralSV analyses. (Observation O7).

(N5) NOT VIABLE - Low confidence predictions of deleterious codons are assigned to codons that can be constructed from observed position frequency matrix from the expanded library, but corresponding AAs are not observed in synonymous codons nor results from SeqalSV/StralSV analyses. (Observation O5 and NOT O7).

(N6) NOT VIABLE - Medium confidence predictions of deleterious codons are assigned to codons that could be constructed based on the expanded position frequency matrix (additional bases predicted from AA positions using approach (V4)), but still not observed in AA results from synonymous codons nor SeqalSV/StralSV analyses. (Observation O8 and NOT O7).

(N7) NOT VIABLE - High confidence predictions of deleterious codons are assigned to all remaining mutation positions based on no evidence in analyzed data that such a mutation or similar one (in homologous proteins) can occur and generate a viable virus. (Observation NOT O8 - no evidence for possible codon construction using data from current genomic or protein sequences).

In assessing probabilities of possible point mutations in the DENV-2 genome we put the highest confidence weights to the predictions that are based on sequence variabilities observed at the nucleotide level within identified DENV-2 genome sequences. Relatively high confidence predictions of new possible mutation points are assigned to the mutations predicted based on calculated alignments between genomes from different Dengue serotypes (DENV-1, -2, -3, and -4). Medium and low confidence predictions are assigned to the predictions, which are based on observed residue-residue correspondences between identified homologous proteins. However, those particular predictions may significantly help identify sequence positions where novel (not observed yet in existing genomic sequence databases) mutations can arise. A critical requirement for a novel codon mutation (e.g. based on observations: O1, O3, O5, O8) to be considered as viable is that the amino acid specified by mutated codon should be observed at the corresponding codon position in genomic sequences from closely related organisms, or in protein sequences identified by protein sequence or structure homology searches (see description of scores: V3, V4, N5, and N6). Finally, mutations for which there is no evidence for existence or theoretical construction using current genomic and protein sequences are assessed as possibly non-viable. In the GeneSV approach, all predictions for the coding regions are evaluated using both sequence and structure similarity analyses at the protein level.

The probability that two given mutation positions are compensating mutations (i.e. mutations whose occurrence at the same time reduces the negative consequences of each of these mutations alone; [40],[41]) is assessed by checking the following characteristics:

- each one of two given mutations should pass a viability check: V1-V4
- using the constructed structural model it is observed that the mutation positions are located in close vicinity to each other (evidence for possible residue-residue interaction, or compensating event when considered mutations share similar physicochemical properties),

- results from calculated multiple sequence alignments show correlation between mutation positions (e.g. switched positions in corresponding proteins from closely related organisms),
- corresponding positions in homologous proteins (identified through sequence or structure homology searches) are annotated as correlated (if information is available, e.g. through literature searches).

In the next two subsections, we describe additional features that are calculated within the GeneSV system to help sequence and structure-based characterization of codon and base positions in the genomic sequences evaluated. Calculated features and their scores are not directly included to the scoring scheme V1-N7, but are reported as additional information in the output from the system to help "manual" refinement of predictions made and to support annotation efforts. Some examples of how these additional characteristics were used to select mutants for experiments performed are provided in subsection 6 below.

4. Assessment of possible position changes by analysis of theoretical probability of codon mutation

The theoretical distance between current and mutated codons can be calculated using the number of transitions and transversions required to generate the mutant codon from the current sequence. Given a sequence with a codon $N=n_1n_2n_3$ in a given coding region position, the probability that a mutant sequence will have a codon $M=m_1m_2m_3$ at the same position, depends on several factors like: the number or replications *n* that led from current sequence to mutant sequence, the various paths in codon space via which N can mutate to M in the course of *n* replications, the 64x64 probabilities P_{ij}=Prob(codon i mutates to codon j during one genome replication), and the probabilities p_i (i=1,…, 64) of each of these codons to generate a viable mutation.

The probabilities Prob(n_i->m_i) depend on whether the mutation n_i -> m_i is a transition ("ts": of type A⇔G or T⇔C) or transversion ("tv": of type A⇔C, G⇔C, A⇔T, G⇔T). For example, assuming that the mutation rate (i.e. probability of a nucleotide change per one genome replication) is q (for RNA viruses it is usually of the order 10^{-3}-10^{-5}, e.g. [25]), and that the probability of transition is twice as large as the probability of transversion, then, in one genome replication the probability of two transversions and one transition in the same codon is proportional to $q^3(1/2)(1/4)(1/4)$, while the probability of one transition only is proportional to q(1/2), with the first probability being 6 to 10 orders of magnitude smaller than the second. In a quasispecies population, high probability mutants (such as single point mutations) of the dominant genotype are expected to be observed with high frequency, unless they produce non-functional genotypes. On the other extreme, mutants of the dominant that would theoretically occur with low probability (such as two or more transitions or transversions in the same codon) but are observed in samples of the quasispecies, must confer competitive advantages.

As an illustration, let ACG be a codon at a given position of some genome sequence in the analyzed quasispecies population. The probability of the codon mutation ACG->GGC is very low (one transition and two transversions = $q^3/32$), so it might be expected that such a mutation at the same position will be observed at a very low frequency in the population. Suppose that GGC has not (yet) been reported/observed in the viral species under study at a specific site in the genome, but has been observed at a similar site in another species whose genome has a high sequence similarity to DENV-2 genome. It is then plausible to assume that GGC might be a viable mutation in DENV-2 but has not been yet observed due to its low probability of mutation from the sequences identified within a quasispecies cloud.

On the contrary, let us assume that the considered mutation codon GGC is only one transition away from the codon AGC observed in the dominant sequence. Then, such a mutation would be expected to occur relatively frequently, and would be expected to be observed. Yet, if this codon has not been observed/reported in databases, even for other related species, it is plausible to assume that this codon would represent a deleterious mutation.

Thus, when assessing probabilities of viable or non-viable mutations for a given clone, we need to refer possible codon changes not only to the positions in the initial (starting) sequence, but also to the positions observed within a cloud (e.g. represented by the dominant sequence), and combine theoretical scores (mutation rates and distances) with estimates that were calculated based on observed/reported corresponding nucleic acid or amino acid positions in other genomic sequences or homologous proteins. For example, to estimate the frequency (and thus the observability) in the DENV-2 quasispecies population of the proposed codon mutations in a mutant sequence of the infectious clone P8-1407 (FLIC), we calculated both distances (in number of transitions and transversions) between: (1) the corresponding codons of the FLIC and mutant sequence, and (2) the corresponding codons of the DENV-2 consensus sequence (RefSeq) and mutant sequence. Obviously, calculated distances are the same if the corresponding codons of the FLIC and RefSeq are identical.

5. Characterization of codon mutations based on the analysis of protein sequences and structures

On the protein level we evaluate the given codon mutation by calculating similarities in sequence and structure context between corresponding proteins from related organisms. For this task, we construct structural models using AS2TS – a homology-based protein structure modeling system, perform sequence-based and structure-based homology searches, and use a sequence variability analysis (SeqalSV), and sequence-structure variability analysis (StralSV) to calculate residue-residue correspondences between identified homologous proteins. To all identified homologous sequences and structures that contribute to the predictions of possible mutations, the following calculated sequence identity and structure similarity scores are assigned:

- sequence identity (Seq_ID) – a sequence similarity score between compared proteins are reported from calculated sequence alignments (e.g. by using Smith-Waterman algorithm) or structure alignments by LGA structure alignment program (when protein structures are available),
- structure similarity (LGA_S) - local and global structural similarity scores between compared protein structures are reported from LGA structure alignment calculations.

We characterize the location of the evaluated mutation position in protein sequence and structure (e.g.: buried, exposed, secondary structure element, conserved region, part of the epitope). We calculate solvent accessibility scores (ACC) and secondary structure assignments (SSE) using the DSSP [26] and STRIDE [27] programs when a reliable structural model is constructed, otherwise the sequence-based prediction methods SSPRO [28] and PSIPRED [29] are used to calculate these features. In the current GeneSV development the epitope predictions (EPI) are performed using the linear B-cell epitope prediction program Bepipred [30]. The following characteristics are assigned to the amino acid positions evaluated:

- ACC(0,1) - buried (0.0<=ACC<=0.20), exposed to the surface (0.20<ACC),
- SSE(C,E,H) – a secondary structure element: C - coil, E - strand, H - helix,
- EPI(1,0) – a position predicted as a part of the antigenic determinant.

In order to characterize possible amino acid substitutions from the sequence conservation point of view we use standard frequency calculations (FR). The estimation of sequence and structure conservation of selected protein regions is performed using constructed structural models and identified protein sequence and structure homologs. Within the GeneSV system, we have implemented two measures: (sCON) and (eCON) that explore well-known strategies for sequence conservation calculations. Let $f_a(i)$ denote a frequency (FR) calculated as an abundance of a given amino acid *a* at the corresponding to *i* positions in closely related organisms, then:

- sCON - sequence similarity conservation using Sum-of-Pairs algorithm is calculated by the formula:

  $$sCON_a(i) = K_s \sum_{b=1}^{21} f_b(i) S_{ab}$$

  where $\{S_{ab}\}$ is an amino acid substitution matrix (e.g. BLOSUM62). In our implementation, in addition to 20 amino acids we use *X* as a 21st character to represent non-standard amino acids. The scoring matrix $\{S_{ab}\}$ is normalized and the constant K_s introduced to produce similarity conservation index from the range [-9.9 , 9.9] (“not similar” to “similar”).

- eCON – estimated amino acid conservation using Shannon's entropy and frequencies is calculated by the SeqalSV and StralSV algorithms. In the implemented formula we use entropy (a measure of uncertainty) with the reverse sign and normalized by constant K_e to produce sequence conservation index from the range [-9.9 , 9.9] (“not conserved” to “conserved”):

$$eCON(i) = 9.9 - K_e \sum_{a=1}^{21} f_a(i) \ln f_a(i)$$

These two measures, sCON and eCON complement each other in the sense that the first one uses information about similarities among amino acids, while the entropy evaluates the overall distribution of observed amino acid frequencies.

6. Selection of position mutations for the GeneSV characterization and experimental validation

Here, we describe the method how a set of mutation points for our study was selected, and illustrate the GeneSV approach by applying it to characterize selected positions and theoretical mutations in the NS5 Polymerase gene from the DENV-2 FLIC. A structural model of NS5 was generated by homology-based modeling system AS2TS using two complementing PDB structures as templates: 3EVG – a crystal structure of the N-terminal domain of DENV-2 virus methyltransferase complexed with s-adenosyl-l-homocysteine (Resolution: 2.20 Å) [31], and 2J7U – a crystal structure of the dengue virus RNA-dependent RNA polymerase catalytic domain solved at 1.85 Ångstrom resolution [32]. The modeling was performed automatically using structural data from PDB chains 3evg_A and 2j7u_A. Side-chain atoms construction was accomplished using SCWRL algorithm [33] when residue-residue correspondences between template and FLIC did not match. Atom positions from residues that were identical in the FLIC and template were copied from the template onto the model. The model was finished with relaxation *via* UCSF Chimera [34]. The constructed model was then used for structure-based homology searches, to perform sequence variability analysis (StralSV), and to structurally characterize candidate positions for mutations.

Our aim was to select a set of diverse positions to test the prediction capabilities and assessment criteria implemented within the developed system. A set of selected positions with amino acids numbers is shown in Figure 1.

Some structural characteristics of this set of 10 selected amino acid positions are described in Table 1. For instance, positions C91 and C179 are buried within the N-terminal domain of the protein while G226 is assessed as exposed to the surface. Positions K279, C400, R437, F483 and K552 are buried. Position W700 is exposed to the tunnel, and position H711 is buried in the tunnel vicinity. In Table 1, we provide: secondary structure element assignments (SSE), solvent accessibility (ACC) scores, predictions if a given position is a part of the epitope (EPI), and results from the conservation calculations (eCON), and (sCON) for FLIC and RefSeq.

For each of the 10 selected positions, we proposed several codon mutations for experimental validation using our reverse genetic system. Our choice of positions and codons to mutate was dictated by the presence of combinations of characteristics in the categories of criteria described above. The goal at this stage was to select a set of positions with diverse characteristics in terms of codon distance, sequence and structure variability, to test the prediction capabilities and assessment criteria, and learn from the validity of made predictions.

For position C91, which is completely buried within the N-terminal domain of RdRp, we expected two mutations with high and low probability to be viable: C91M - 3 transversions away, and C91A - 2 transversions away from the corresponding codon in RefSeq from DENV-2. The basis for these predictions was that Methionine and Alanine have been observed in corresponding positions in related organisms (M91 in DENV-4 and A91 in Murray Valley encephalitis virus) with fairly high sequence identity 76.0% and 66.7%, respectively (Table 3). One of the main differences in our confidence assigned to those predictions is that the codon ATG (defining M) is observed at the corresponding position in DENV-4 while the codon GCT (defining A) cannot be even constructed from the calculated DENV-1234 position frequency matrix. We predicted these mutations to be viable based on the criteria V2 and V4, and this assessment was experimentally validated.

Two mutations (C179A and C179V) proposed at position C179 had the same relatively high codon distance of 2 transversions. As shown in Table 1, the Cysteine at this buried position located within a strand region (E) is highly conserved (8.9). This suggests low sequence diversity that could be expected at this position, which is confirmed by calculated sCON estimates. For the proposed mutations A and V, the sequence conservation scores sCON are as low as: -2.0 and -4.0, respectively (see Table 3). However, these two small hydrophobic amino acids (A and V) had been observed at the corresponding positions in other organisms (179A – Iguape virus; 179V – Barkedji virus) with moderately high sequence identity of 63.4% and 62.6%, respectively. Therefore, in this experiment we wanted to test the dependence on these assigned characteristics. Our predictions of these mutations to be viable were based on the criterion O7 with confidence V4 and were correct.

For position G226, we suggested a variety of mutations. This position was chosen because it is located in the coil (C) region exposed to the surface, highly conserved with eCON as high as 8.9, and is a part of the predicted antigenic determinant (see Table 1). For this position, we selected three mutations G226N, G226S, and G226T based on moderately high sequence identity to the sequences from other organisms where selected amino acids are observed at corresponding positions: Wesselsbron virus (63.3%), - Rocio virus (68.0%), and 226T - Iguape virus (63.4%). Our predictions for these first three mutations to be viable were based on the criterion O7 with confidence V4. The fourth mutation G226E had opposing characteristics, compared to the previous ones in that the G(GGG)->E(GAG) codon mutation is realized by only one transition (minimum distance, so the mutation seems to be easy to achieve), yet 226E has not been observed in other species, E is a negatively charged amino acid while not charged amino acids are observed at the corresponding positions, and the codon GAG cannot be constructed from calculated DENV-1234 position frequency matrix. Based on criterion O8 and confidence level N6 and the fact that this is an “easy, mutation” yet not observed in any close organism we expected the fourth mutation to be deleterious. All four predictions were positively validated by the experiment (Table 3).

From all positions listed in Table 1, K279 has the lowest estimated amino acid conservation (eCON). So, it would be expected that a diverse list of possible substitutions can be allowed at this position.

Interestingly, we found via GeneSV searches that the clone gi|253967733|dbj|DM140335.1 from a DENV-2 attenuated strain has an Arginine defined by the codon AGA at the position corresponding to K279 (8404: K(AAA)). We further found two other amino acids observed in corresponding positions in related organisms (Glutamic acid in DENV-3 and Asparagine in West Nile virus). For the experiment, we proposed at this position two mutations: 279E and 279N, and both were confirmed as viable mutations (as predicted with confidences V2 and V4, correspondingly). Taking into account the possible importance of this position and mutation K279R for virus attenuation, we also selected this position as the part of the evaluated compensating mutation pair (**K279** and **R437**) described in Table 4.

We chose position C400 (with C codon TGC) as especially interesting in that the infectious clone differs in this position in all three nucleic bases (three transversions) from the RefSeq (which has Threonine coded by ACA at position 400) and that this is the only triple transversions mutation (referring to RefSeq) in the entire sequence of the infectious clone (FLIC). For this position we proposed two experiments: a codon mutation to Threonine created by three transversions (ACA), and codon mutation to Threonine by two transversions (ACC). By these experiments we wanted to check if such a maximum distance (three transversions) single codon change could still produce a viable virus or it would requires some other assisting or intermediate mutations. Based on our current criteria proposed mutations were assigned with the categories O0 and O1-O5 and confidence V1 and V3, respectively. It was confirmed by the experiment that they produced viable viruses (see Table 3).

For position R437, which shows medium eCON conservation (2.9) we proposed a diverse set of mutations 437A, 437Q, 437S, and 437E to test the confidence levels V4 and V3 in our predictions. Based on the presence of listed amino acids at the corresponding positions in sequences from other organisms and observed high sequence identities we were fairly confident in the viability of our proposed mutations. These predictions were confirmed to be correct (see Table 3).

For position F483 located in relatively conserved region (eCON 5.9), buried, and in the helical region, we proposed 4 codon mutations: 483S, 483Y, 483V, and 483L. In each case the codon distance from the mutant to FLIC was small (one transition or one transversion). Assessment of the mutation 483S (with S codon TCC) showed that S had not been observed in other species, nor could be constructed from all possible bases that could be observed at corresponding position in other species (i.e. criterion O8 is not fulfilled for this mutation). We predicted with confidence N7 that this mutant will be deleterious, and this prediction was correct. In the case of the mutation 483Y the GeneSV system reported a genomic sequence in DENV-3 with Tyrosine at the corresponding position and high sequence identity of 85.4% to FLIC, so with confidence V2 we expected this mutation to be a viable one. 483V and 483L have similar to each other characteristics: low sCON conservations (-4.0 and -2.0), and very low sequence identity (20.3% and 17.9%) to proteins from other organisms (HCV and FMDV, respectively) where V and L could be observed at a given positions. We felt less confident in the viability of 483V and 483L because of identified very low sequence identities. Yet, based on the assigned categories O4 and O7 we predicted with low confidence that these mutations may be viable. These two predictions were demonstrated to be incorrect (see Table 3). A lesson learned from these predictions is that when a very few supporting sequences can be identified (see results marked by '*' in Table 3), and the level of sequence identities to the supporting sequences is very low, then such evidence for possible mutations should be treated as unreliable.

For position K552 (helical, buried, and part of predicted epitope peptide) with medium eCON conservation of 1.9 we proposed 4 codon mutations: 552L, 552Q, 552I and 552N. Each prediction was supported by the presence of the listed amino acids at the corresponding positions in the sequences from virus organisms other than DENV. The levels of sequence identities to the identified protein sequences varied from 66.6% to 92.9%. The confidence levels of evaluated mutations varied from V2 to V4, and all tested mutations were experimentally confirmed as viable.

Position W700 was chosen as an example of a position exposed to the tunnel, and also because Tryptophan (coded by TGG) is known as an amino acid which rarely mutates (results from the conservation analysis showed at this position 6.9 of eCON and 8.9 of sCON for W). Six mutant codons were proposed, 4 of which were at either two transversions (F(TTC) and A(GCG)) or 1 transversion and 1 transition (Y(TAC) and E(GAG)) away from W(TGG) from FLIC. One mutant D(GAC) is one transition and 2 transversions, and one C(TGC) is only one transversion away from Tryptophan. Based on the results from sequence similarity searches we had a fairly good support for assigning V4 confidence to the prediction that mutations 700Y and 700F will be viable. We had less convincing support for predicting viability of the mutation 700A because of the low sequence identity (18.5%) to the closest identified related protein sequence. Nevertheless, all three predictions were experimentally confirmed as correct. Similarly to 226E (above), mutant with 700D had not been observed in other species, D is a negatively charged amino acid while not charged amino acids are observed at the corresponding positions, and the codon GAC cannot be constructed from calculated DENV-1234 position frequency matrix (see Table 3). Based on criterion O8 (confidence N6) we expected this mutation to be deleterious, and we were correct. Similar arguments were used for assigning N6 confidence to mutation 700C (except that Cysteine is polar but neutral amino acid). The experiment showed that our assessment for this mutation was wrong. Although our confidence in the remaining prediction (700E) was rather low, the prediction was made with the goal to learn from the experimental validation. We predicted this mutation as viable with confidence V4 based on O7 evidence despite E being a charged residue while W700 is assessed as conserved neutral position. The experiment showed that mutation 700E produced not viable mutant.

Position F711 was chosen for testing two predictions in a region buried in the tunnel vicinity and assessed as conserved (eCON 5.9) with dominating charged residues (see '7' at 'ch' column from Table 3). In this case, our V4 confidence predictions 711N of neutral mutation in the dominated by charged residues position was incorrect. In our second prediction for this position we expected mutation 711R (1 transition away charged substitution in the charged position) to be not viable and, we were again wrong.

Detailed results from the GeneSV analysis of discussed above 32 single codon mutations are provided in the supplemental data – Suppl 1.

Two of the selected above single mutation positions **K279** and **R437** were also proposed for testing of possible compensating mutations. Our higher confidence prediction was:

If there is R or K in one of these positions (no mutation, or switch: **279R** or **437K**),

then a change to E in the other (**279E** or **437E**) should produce VIABLE mutant.

If both positions are changed to E (**279E** and **437E**), then it may produce NOT VIABLE mutant.

Our low confidence prediction was:

If there is N or Q in one of these positions (e.g.: **279N** or **279Q** or **437N** or **437Q**),

then a change to E in the other (**279E** or **437E**) should result in VIABLE mutant.

If both positions are changed to E (**279E** and **437E**), then it may produce NOT VIABLE mutant.

Results from the experimental validations of above predictions of possible compensating mutations are provided in Table 4.

7. Experimental generation of DENV-2 mutants

Plasmids containing point mutations were generated from DENV-2 cDNA clone [23]. Briefly, the DENV-2 genome is expressed from a cytomegalovirus (CMV) promoter. Stability of the cloned genome is ensured by the incorporation of a stabilizing intron sequence engineered at the junction of envelope (E) and the non-structural 1 (NS1) gene. A unique feature of the infectious clone is the insertion of the hepatitis Delta virus ribozyme (HDVr) immediately after the last nucleotide of DENV-2 cDNA sequence to ensure production of DENV RNAs with the precise, correct 3'- terminus, which is beneficial for more efficient RNA replication. Furthermore, a SV40 polyadenylation site was inserted downstream of the HDVr to ensure complete termination of transcription. Mutations in the polymerase gene (nonstructural gene 5 (NS5)) were inserted into specific positions in the genome using standard recombinant DNA techniques described elsewhere [35]. Rescue of the generated mutants was achieved by transfecting 4 μg of the respective plasmids into Vero E6 (African green monkey kidney) cells seeded in 6-well plates at $5x10^5$ cells/well using Lipofectamine 2000 (Invitrogen, Grand Island, NY) according to manufacturer's protocol. Supernatants containing rescued viruses were harvested 5 days later and their potency determined by focus forming immunoassay (FFA) as described previously [36].

RESULTS

We performed two independent experiments to assess predictive accuracy of the GeneSV system. For the first experiment, we used two complete datasets of DENV genomic sequences downloaded from public databases (GenBank and ENA). The first dataset was downloaded on February 12th, 2012, and second downloaded on February 18th, 2013. We used the RefSeq sequence of DENV-2 for mapping mutations and the first dataset as an input library for GeneSV calculations. A set of 1205 novel mutations (1170 new codons in coding regions, and 35 new base positions in noncoding regions) that were identified in the second dataset and not present in the input library, were used as test mutations for GeneSV predictions. In this experiment the system was able to predict 1139 codons (97%) and 29 base positions (83%) as viable mutations. For our second experiment to evaluate the GeneSV system, we chose 10 different codon positions on the NS5 gene (RNA-dependent RNA polymerase (RdRp)). The NS5 gene was taken from a full-length infectious clone (FLIC) of DENV-2 strain P8-1407 (gi:132271146). At each position we made predictions with different confidence levels as to whether a number of chosen amino acid substitutions would produce viable or non-viable virus. We reasoned that a mutant RdRp that is predicted to lack viability would lead to a non-viable virus. The predictions were tested by generation of point mutations in the NS5 gene and attempts to rescue infectious DENV-2 from the mutated infectious clone plasmids on Vero cells [6]. From the selected set of 32 single mutation positions proposed, 25 were experimentally confirmed as viable and 7 as non-viable. The GeneSV system was correct in its assessment in 26 cases (81%). In order to evaluate the robustness of created predictions we performed GeneSV calculations twice. One prediction was made using first dataset collected on February 12th, 2012, and the second prediction was made using a dataset from February 18th, 2013. As we report in Table 3 (columns GSV1 and GSV2) no significant change in created predictions was observed in the results from the two datasets.

1. GeneSV assessment of 1205 mutations taken from updated datasets of genomic sequences

In this test we evaluated a set of 1205 mutations obtained from comparison of two datasets (2012.02.12 and 2013.02.18) of genomic sequences downloaded from public databases which differ by 1338 sequences. The differences were comprised of:

- 538 more sequences for DENV-1 in dataset (2013.02.18) than in dataset (2012.02.12)
- 399 more sequences for DENV-2
- 154 more sequences for DENV-3
- 159 more sequences for DENV-4
- 88 more Dengue sequences with no specified type

The GeneSV analysis showed that within 399 new DENV-2 sequences there are 1170 novel codon positions in the coding regions and 35 new base positions in the non-coding regions (5' and 3' ends). Results from the GeneSV assessment of these 1205 mutations are reported in Table 2 below.

Results from the GeneSV assessment of new mutations defined by comparison of two datasets 2012.02.12 and 2013.02.18 showed that from 1205 (1170 codons + 35 bases) new mutations observed in updated DENV-2 genomic sequences 1139 out of 1170 codon mutations (97%), and 29 out of 35 mutations from noncoding regions (83%) were predicted correctly as viable. Detailed results from the GeneSV analysis of these mutations are provided in the supplemental data – Suppl 2.

2. Frequencies of observed codon variants in DENV-2 protein coding regions

GeneSV can be used for various purposes aimed at characterizing sequence variability. For example, we applied GeneSV to estimate frequencies of mutations observed in codon positions within different genes within the set of all available genomic sequences (dataset 2013.02.18). Of course, the number of different nucleotides at a given position varies from 1 to 4, and the number of possible codons at each protein coding region position can be as low as 1 or as high (theoretically) as 64. In Figure 2, we report calculated frequencies of observed mutations (number of codon variants) with respect to their locations specified by different amino-acid and position characteristics within each protein coding region of the DENV-2 genome: secondary structure elements (coil, strand, helix), solvent accessibility (buried, exposed), and regions predicted as potential antigenic determinants (not epitope, epitope).

Reported results show that for the DENV-2 genome, the most mutable regions are observed in the Envelope (Obsrvd: 5.19) within the segments characterized as helical (5.46), exposed (5.27), and predicted as antigenic determinants (5.22). On average, for all coding regions (see AVERAGE set of bars in Figure 2) those frequencies are: 3.93, 3.99, 3.84, and 3.78, respectively. The lowest mutability is observed in NS3 protease (Obsrvd: 3.16). Results show that high mutabilities, above the genome average, are seen in the structural proteins: Capsid, prM, and Envelope. Interestingly, two non-structural proteins Nsp-1 and Nsp-2A are also characterized by relatively high mutability. Detailed results from the GeneSV analysis are provided in the supplemental data – Suppl 3.

3. GeneSV assessment and experimental validation of 32 hypothetical mutations selected in RdRp

To experimentally test the accuracy of GeneSV in assessing possible viability of codon mutations, we selected in the gene coding RdRP protein 10 different codon positions for making predictions. Selected codon positions are described in Table 1 and their location on the structural model of RdRp is shown in Figure 1. For each position several codon mutations have been proposed, and assessment generated by the GeneSV system was compared with experimental results. In Table 3 we list 32 single codon mutations (not seen in DENV-2 sequence data; except C400) that were selected for this study. Attempts were made to rescue these 32 variants using the DENV-2 FLIC. In the assessment of the proposed mutations, based on the criteria described in the Materials and Methods section, we expected the mutation predictions characterized by confidence levels (V1) - (V4) to be viable, while mutations with confidence levels (N5) - (N7) to be deleterious. The experiment showed that in 7 cases the introduced codon mutations (colored in blue in Table 3) did not yield viable viruses. Results from Table 3 show that the proposed approach to assess the viability of given mutations was correct in 26 out of 32 cases (81%) as confirmed by the experiment. Only 6 predictions were incorrect. Wrong predictions are colored in red, and correct predictions are colored in green. The last four columns in Table 3 show comparison of the viability predictions by different methods: “GSV” – viability prediction by GeneSV system (GSV1 – predictions based on the dataset 2012.02.12, GSV2 – predictions based on the dataset 2013.02.18), SIFT – predictions by SIFT algorithm [14], and PRV – predictions by PROVEAN [15].

In Table 4, we show a list of 5 double mutations that used positions K279 and R437 to test GeneSV predictions for potential compensatory mutations, with four out of the five paired substitutions correctly predicted as viable/non-viable.

The predictions by SIFT and PROVEAN reported in Table 3 were created using publicly available web services. Results show that using their default settings the accuracy of SIFT, PROVEAN and GeneSV is comparable, ranging from 78% to 81% and from 0.837 to 0.885 by F1 scores (see last column in the list below.)

PROVEAN	http://provean.jcvi.org/seq_submit.php	7/32 (78%; 0.837)
SIFT	http://sift.jcvi.org/www/SIFT_seq_submit2.html	7/32 (78%; 0.857)
GeneSV	http://as2ts.llnl.gov/GENESV/	6/32 (81%; 0.885)
Prediction by consensus of SIFT, PROVEAN, and GeneSV		5/32 (84%; 0.902)

In Table 3 and in the list above, the number of wrong predictions created by a given method is highlighted in red. It shows, for example, that in our tests the SIFT method was wrong in predicting 3 non-viable and 4 viable mutations while the GeneSV was wrong in 4 non-viable and 2 viable mutations. Interestingly, the accuracy score of 78% achieved by the PROVEAN method in our experiment agreed well with estimations reported by PROVEAN’s authors (77% - [15]) when the method was applied to different sets of protein variants from viruses, fungi, bacteria or plants. It is worth to notice that among the 7 non-viable mutants in our tests, PROVEAN was correct in all seven cases, but it failed in predicting 7 viable mutations - probably due to being too caution in predicting possible viability when no strong evidence of conservation or similar residues at a given position can be detected among homologous proteins. On the other hand, as we discuss in the section below, in the case of RNA viruses results from site-directed mutagenesis experiments suggest that the number of random single nucleotide mutations in a given genome (e.g. Vesicular Stomatitis Virus – [38]) identified as viable can be as high as 60%. It means that in some cases even a trivial prediction approach which assesses all mutants in performed experiments as viable can achieve a high proportion of correct predictions. We showed that a combination of information (e.g. like the one implemented within the GeneSV system) about the permissible nucleotide variability observed in corresponding regions in genomic sequences, with conservation/variability results from protein sequence and structure-based analyses can improve overall accuracy predictions, i.e. in both viable and non-viable cases. Moreover, when the predictions from all three services are combined, then the achieved accuracy can be even higher. A brief summary of similarities and differences between homology searches and conservation analysis approaches implemented within evaluated mutation analysis tools is provided in Table 5.

DISCUSSION

GeneSV is a computational system designed to facilitate assessment of regions of sequence variability in genomic sequences. Such assessments may be very valuable especially for RNA viruses, which are characterized by high mutation rates and presence in highly diverse populations [37]. For its analysis, the system combines information from a large variety of sources about the permissible nucleic acids as well as amino acid variabilities calculated from evaluated non-coding regions or protein-coding genes from collected genomes. In addition to genetic information, the system takes into consideration the primary protein sequences coded for, and structure-based analyses of the resultant tertiary structures. Generated results can help functional annotation of evaluated genomes, predict potential nucleic- and amino acid mutations not observed in current databases, or assess the accuracy of novel mutation positions reported from the sequencing efforts. In this manuscript we showed how the GeneSV system was applied to predict the functional effects of amino acid substitutions in 10 positions selected on the RNA-dependant RNA polymerase (RdRp) of Dengue virus type 2 (DENV-2). At each position we made predictions with different confidence levels as to whether a number of chosen amino acid substitutions would produce viable or non-viable virus. We evaluated predictions of 32 single amino acid substitutions, 31 of which had never been observed in any publicly available DENV-2 sequence. In 81% of the predictions (26 of 32), GeneSV predicted the correct phenotype: functional vs. non-functional RdRp as measured experimentally by rescue of virus using a DENV-2 clone. Five additional mutants with double amino acid substitutions proximal in structure to each other were generated and in 80% of cases (4 of 5) GeneSV was correct in its predictions. For these predictions we performed GeneSV calculations twice. One prediction was made using a dataset collected on February 12th, 2012, and the second prediction was made using a datasets from February 18th, 2013. The predictions, GSV1 and GSV2 (see Table 3), were almost identical showing that results from GeneSV were consistent and did not change significantly with the growth of the databases.

In another experiment, a set of 1205 novel mutation positions that were identified from updated databases was used for viability assessment. Results from that experiment showed that the system was able to correctly characterize as viable mutations 97% codon positions from coding regions and 83% base positions from noncoding regions.

Another interesting and practical question is how to estimate the size of the virus quasispecies cloud. The GeneSV system would enable such estimation if a library of genomic sequences adequate for the specific study were provided (e.g. library which would be representative for a given environment like geographic area, hosts, etc.). For example, for protein coding regions from FLIC, with 3391 amino acids in size, the number of all theoretically possible codons is 217,024=64*3391 (64 mutations per each position). Now, if all available DENV-2 sequences from dataset 2013.02.18 are used for calculations, the following estimates can be generated by the current version of the GeneSV system: 44% (94,929) of theoretically possible codons predicted as viable, and 56% (122,095) predicted as non-viable. This estimate of possible viable codons seems rather high, however, in some publications based on site-directed mutagenesis experiments it is suggested that the number of possible random single nucleotide mutations in a given genome of an RNA virus (e.g. Vesicular Stomatitis Virus – [38]) can be even higher - as high as 60% [39]. Our estimate of 44% suggests that on average for each codon position up to 28 different mutations could produce viable DENV-2 mutants (codon mutability). On the other hand, based on the current observation, which will inevitably change by growing numbers in the future, the estimate of average DENV-2 codon mutability is 3.93 (Figure 2) and we could expect that the true value is somewhere between this number and the one predicted above. We believe that if we use libraries limited to concrete environmental characteristics (not just all available sequences from different sequencing projects), our predictions will become more accurate and closer to reality.

Moreover, as we showed in the analysis of the results from the characterization of selected 32 codon mutations, the predictive accuracy of the system can be improved by better selection and weighing of calculated scores. Additional improvements can be achieved when refined and organism specific libraries are used for final assessment of possible viability of evaluated mutations in the proteins from a given species. Further study of this topic will require access to more complete results from the experiments when sequences with confirmed not viable mutations are also reported and in the well-defined computer friendly format (current sequence databases mostly focus on providing positive results from the sequencing efforts without easy to grasp additional details).

SUPPLEMENTAL DATA

Supplementary Data is available online and includes the following results:

1. Summary output from the GeneSV assessment of selected 32 single codon hypothetical mutations in DENV-2 infectious clone P8_1407 (FLIC) using all collected DENV-1234 genomic sequences 2012.02.12:
   http://as2ts.llnl.gov/AS2TS/GENESV/DOCS/Summary_GeneSV.P8_1407_mutants.prediction_20120212.UTMB_validated

2. Summary output from the GeneSV assessment of new 1205 mutations in DENV-2 RefSeq not seen in DENV-2 genomic sequences 2012.02.12 using all collected DENV-1234 genomic sequences 2013.02.18:
   http://as2ts.llnl.gov/AS2TS/GENESV/DOCS/Summary_GeneSV.RefSeq.update_20120212

3. Summary output from the GeneSV assessment of all possible single codon mutations in DENV-2 infectious clone P8-1407 (FLIC) using all collected DENV-1234 genomic sequences 2013.02.18:
   http://as2ts.llnl.gov/AS2TS/GENESV/DOCS/Summary_GeneSV.P8_1407.status_20130218.all_64

A fully automated version of the GeneSV system is under construction:
http://as2ts.llnl.gov/AS2TS/GENESV/

ACKNOWLEDGMENT

Prepared by LLNL under Contract DE-AC52-07NA27344.

FUNDING

This work was funded by the United States Department of Defense, Defense Advanced Research Projects Agency's Prophecy Program: Pathogen Defeat.

REFERENCES

[1] Drake D.W., Holland J.J. (1999) Mutation rates among RNA viruses, Proc. Natl. Acad. Sci. USA, 96, pp. 13910–13913

[2] Duffy S., Shackelton L.A., Holmes E.C. (2008) Rates of evolutionary change in viruses: Patterns and determinants. Nat Rev Genet 9:267–276

[3] Duarte E.A., Novella I.S., Weaver S.C., Domingo E., Wain-Hobson S., Clarke D.K., Moya A., Elena S.F., De La Torre J.C., Holland J.J. (1994) RNA virus quasispecies: significance for viral disease and epidemiology. Infect. Agents Dis. 3:201–214.

[4] Domingo E., Biebricher C., Eigen M., Holland J.J. (2001) Quasispecies and RNA virus evolution: principles and consequences. Landes Bioscience, Austin, Texas

[5] Vignuzzi M., Stone J.K., Arnold J.J., Cameron C.E., Andino R. (2006) Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 439: 344–348.

[6] Holland J., Spindler K., Horodyski F., Grabau E., Nichol S., VandePol S. (1982) Rapid evolution of RNA genomes. Science.215:1577–85.

[7] Holland J.J., De La Torre J.C., Steinhauer D.A. (1992) RNA virus populations as quasispecies. Curr Top Microbiol Immunol.176:1–20.

[8] Domingo E., Holland J.J. (1997) RNA virus mutations and fitness for survival. Annu Rev Microbiol.51:151–78.

[9] Benson D.A., Cavanaugh M., Clark K., Karsch-Mizrachi I., Lipman D.J., Ostell J., Sayers E.W. (2013) GenBank. Nucleic Acids Res. 41(Database issue):D36-42. doi: 10.1093/nar/gks1195.

[10] Leinonen R., Akhtar R., Birney E., Bower L., Cerdeno-Tarraga A., Cheng Y., Cleland I., Faruque N., Goodgame N., Gibson R., Hoad G., Jang M., Pakseresht N., Plaister S., Radhakrishnan R., Reddy K., Sobhany S., Ten-Hoopen P., Vaughan R., Zalunin V., Cochrane G. (2011) The European Nucleotide Archive. Nucleic Acids Res. 39:D28-D31.

[11] The UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt), Nucleic Acids Res. 40: D71-D75.

[12] Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. (2000) The protein data bank. Nucleic Acids Res. 8, 235-242.

[13] Cooper G.M., Shendure J. (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 12: 628–640.

[14] Kumar P., Henikoff S., Ng P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm, Nature Protocols, 4(8): 1073-1082

[15] Choi Y., Sims G.E., Murphy S., Miller J.R., Chan A.P. (2012) Predicting the Functional Effect of Amino Acid Substitutions and Indels", PLoS ONE, 7(10): e46688

[16] Adzhubei I.A., Schmidt S., Peshkin L., Ramensky V.E., Gerasimova A., Bork P., Kondrashov A.S., Sunyaev S.R. (2010) A method and server for predicting damaging missense mutations, Nat Methods, 7(4): 248–249

[17] Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17), 3389-3402.

[18] Katoh K., Misawa K., Kuma K., Miyata T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 15;30(14):3059-66.

[19] Pearson W.R. (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics,11:635-650.

[20] Zemla A., Zhou C.E., Slezak T., Kuczmarski T., Rama D., Torres C., Sawicka D., Barsky D. (2005) AS2TS system for protein structure modeling and analysis, Nucleic Acids Res. 33, pp. W111-W115.

[21] Zemla A., Lang D.M., Kostova T., Andino R., Zhou C.L. (2011) StralSV: assessment of sequence variability within similar 3D structures and application to polio RNA-dependent RNA polymerase, BMC Bioinformatics, 12(1):226.

[22] Zemla A. (2003) LGA - a method for finding 3D similarities in protein structures, Nucleic Acids Res. Vol. 31, No. 13, pp. 3370-3374.

[23] Tsetsarkin K., Widen S., Wood T.G., Hanley K.A., Allen J., Naraghi-Arani P., Weaver S.C., Vasilakis N. Development and Characterization of a Stable Reverse Genetics System of a Malaysian Sylvatic Dengue Virus Type 2 Strain (P8-1407), manuscript in preparation

[24] Rudnick A., Lim T.W. (1986) Dengue fever studies in Malaysia. Institute of Medical Research of Malaysia Bulletin 23:51-152

[25] Domingo E. (2007) Virus Evolution, in: Fields Virology, Fifth Edition, Knipe D.M. and Howley P.M. (Eds), Walter Kluwer/Lippincott, Williams and Wilkins

[26] Kabsch W., Sander C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22, 2577-2637.

[27] Frishman D., Argos P. (1995) Knowledge-based protein secondary structure assignment. *Proteins* 23(4):566-79.

[28] Cheng J., Randall A.Z., Sweredoski M.J., Baldi P. (2005) SCRATCH: a Protein Structure and Structural Feature Prediction Server, Nucleic Acids Res. Web Server Issue, vol. 33, 72-76

[29] Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292: 195-202.

[30] Larsen J., Lund O., Nielsen M. (2006) Improved method for predicting linear B-cell epitopes. Immun Res. 2:2.

[31] Geiss B.J., Thompson A.A., Andrews A.J., Sons R.L., Gari H.H., Keenan S.M., Peersen O.B. (2009) Analysis of flavivirus NS5 methyltransferase cap binding. *J Mol Biol*, **385**, 1643-1654.

[32] Yap T.L., Xu T., Chen Y.L., Malet H., Egloff M.P., Canard B., Vasudevan S.G., Lescar J. (2007) Crystal structure of the dengue virus RNA-dependent RNA polymerase catalytic domain at 1.85-angstrom resolution. *J Virol*, **81**, 4753-4765.

[33] Krivov G.G., Shapovalov M.V., Dunbrack Jr R.L. (2009) Improved prediction of protein side-chain conformations with scwrl4. *Proteins 77*, 4, 778-95.

[34] Pettersen E.F., Goddard T.D., Huang C.C., Couch G.S., Greenblatt D.M., Meng E.C., Ferrin T.E. (2004) Ucsf chimera--a visualization system for exploratory research and analysis. J Comput Chem. 25, 13, 1605-12.

[35] Yamshchikov V.F., Wengler G., Perelygin A.A., Briton M.A., Compans R.W. (2001) An Infectious Clone of the West Nile Flavivirus, Virology 281:294-304.

[36] Vasilakis N., Shell E.J., Fokam E.B., Mason P.W., Hanley K.A., Estes D.M., Weaver S.C. (2007) Potential of ancestral sylvatic dengue-2 viruses to re-emerge. Virology 358 (2), 402–412.

[37] Domingo E., Sheldon J., Perales C. (2012) Viral Quasispecies Evolution, Microbiol. Mol. Biol. Rev. 76(2):159.

[38] Sanjua´n R., Moya A., Elena S.F. (2004) The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc Natl Acad Sci USA 101: 8396–8401.

[39] Lauring A.S., Andino R. (2010) Quasispecies Theory and the Behavior of RNA Viruses, PLoS Pathogens, 6(7), e1001005.

[40] Davis B.H., Poon A.F.Y., Whitlock M.C. (2009) Compensatory mutations are repeatable and clustered within proteins, Proc.R.Soc B, 276: 1823-1827.

[41] Kuipers R.K.P., Joosten H.J., Verwiel E., Paans S., Akerboom J., Oost J., Leferink N.G.H., Berkel W.J.H., Vriend G., Schaap P.J. (2009) Correlated mutation analyses on super-family alignments reveal functionally important residues. Proteins, 76: 608–616.

Residue-position	SSE	ACC	EPI	eCON	sCON (FLIC)	sCON (RefSeq)
C-91	H	0	0	4.9	5.6 (C)	5.6 (C)
C-179	E	0	0	8.9	8.9 (C)	8.9 (C)
G-226	C	1	1	8.9	8.9 (G)	8.9 (G)
K-279	H	0	0	-0.9	2.2 (K)	2.2 (K)
C-400	H	0	0	4.9	-3.0 (C)	4.6 (T)
R-437	H	0	0	2.9	4.5 (R)	2.1 (K)
F-483	H	0	0	5.9	7.7 (F)	7.7 (F)
K-552	H	0	1	1.9	3.5 (K)	-1.2 (M)
W-700	E	1	1	6.9	8.9 (W)	8.9 (W)
H-711	E	0	0	5.9	7.7 (H)	7.7 (H)

Table1: Protein sequence and structure-based characteristics calculated for 10 selected positions in the RNA-dependent RNA polymerase gene of DENV-2 from a full-length infectious clone [FLIC] used to test our predictions. **SSE:** secondary structure element assignments, **ACC:** solvent accessibility scores, **EPI:** predictions that indicate that a given position is a part of the epitope, **eCON:** estimated amino acid conservation using Shannon's entropy and frequencies calculated by the SeqalSV and StralSV algorithms, **sCON:** sequence similarity conservation using Sum-of-Pairs algorithm, **H:** helix, **C:** coil, **E:** strand

Name	Size	New-mutations	Predicted Viable	Predicted Non-Viable
Capsid	114	57	51	6
prM	166	55	50	5
Envelope	495	250	237	13
Nsp_1	352	106	105	1
Nsp_2A	218	129	128	1
NS2B	130	55	55	0
NS3	618	146	146	0
Nsp_4A	150	72	70	2
Nsp_4B	248	68	66	2
NS5	900	232	231	1
TOTAL Codons **(In coding regions)**	**3391**	**1170**	**1088**	**25**
5-end(*)	96*	4*	3*	1*
3-end(*)	456*	31*	26*	5*
Total Bases **(In non-coding regions)**	**552***	**35***	**29***	**6***

Table 2: Results from GeneSV assessment of new mutations identified by comparison of two datasets of genetic sequences separated by 12 months: 2012.02.12 and 2013.02.18. **Name:** region in the DENV-2 genome using RefSeq as a reference sequence. **Size:** number of codons in protein coding regions or bases in noncoding regions (marked by *). **New:** number of new mutations observed in the library 2013.02.18 and NOT seen in 2012.02.12. **Predicted Viable:** number of mutations predicted to be “VIABLE” (correct predictions). **Predicted Non-viable:** number of mutations predicted as “NOT viable” (incorrect predictions)

#Base	Rnum	Cod2	A	TsTv		Cod1	A	1	2	3	4	5	6	7	8	ch	sCON	gcSid4	aaSidH	GSV1	GSV2	SIFT	PRV
7840	91	TGT	C	0	3	ATG	M	1	1	1	1	1	1	1	1	0	-2.1	70.80	76.00	V2	V2	V	V
7840	91	TGT	C	0	2	GCT	A	0	0	0	0	0	0	1	1	0	-1.1	-	66.70	V4	V4	V	V
8104	179	TGC	C	0	2	GTC	V	0	0	0	0	0	0	1	1	0	-4.0	-	62.60	V4	V4	V	N
8104	179	TGC	C	0	2	GCC	A	0	0	0	0	0	0	1	1	0	-2.0	-	63.40	V4	V4	V	N
8245	226	GGG	G	1	1	ACG	T	0	0	0	0	0	0	1	1	0	-8.0	-	63.40	V4	V4	V	V
8245	226	GGG	G	2	1	AAC	N	0	0	0	0	0	0	1	1	0	-1.0	-	63.30	V4	V4	N	V
8245	226	GGG	G	1	0	GAG	E	0	0	0	0	0	0	0	1	+0	-5.0	-	-	N6	N6	N	N
8245	226	GGG	G	1	1	AGC	S	0	0	0	0	0	0	1	1	0	-2.0	-	68.00	V4	V4	V	V
8404	279	AAA	K	0	1	AAC	N	0	1	1	0	0	1	1	1	4	-0.1	-	66.90	V4	V4	V	V
8404	279	AAA	K	1	0	GAA	E	0	0	0	1	1	1	1	1	+4	2.2	73.90	81.00	V2	V2	V	V
8767	400	TGC	C	0	2	ACC	T	1	1	1	1	1	1	1	1	0	4.6	*69.20	93.80	V3	V3	V	V
8767	400	TGC	C	0	3	ACA	T	1	1	1	1	1	1	1	1	0	4.6	82.70	93.80	V1	V1	V	V
8878	437	AGA	R	1	1	GCA	A	0	0	0	0	0	0	1	1	8	-4.0	-	62.40	V4	V4	V	V
8878	437	AGA	R	1	1	CAA	Q	0	0	1	0	0	0	1	1	8	0.0	-	65.50	V4	V3	V	V
8878	437	AGA	R	2	0	GAA	E	1	1	1	1	1	1	1	1	+8	0.2	*67.70	72.80	V3	V4	V	V
8878	437	AGA	R	0	1	AGC	S	0	0	1	0	0	0	1	1	8	-3.0	-	57.10	V4	V4	N	V
9016	483	TTC	F	1	0	CTC	L	0	0	0	0	0	0	1	1	0	-2.0	-	17.90	V4	V4	V	N
9016	483	TTC	F	0	1	GTC	V	0	0	0	1	1	1	1	1	0	-4.0	*72.90	20.30	V3	V4	N	N
9016	483	TTC	F	1	0	TCC	S	0	0	0	0	0	0	0	0	0	-6.0	-	-	N7	N7	N	N
9016	483	TTC	F	0	1	TAC	Y	0	0	0	1	1	1	1	1	0	4.2	85.40	88.90	V2	V2	V	V
9223	552	AAG	K	0	2	CTG	L	1	1	1	1	1	1	1	1	5	-3.0	70.80	92.40	V2	V2	V	V
9223	552	AAG	K	1	1	ATA	I	1	1	1	1	1	1	1	1	5	-5.0	*81.40	92.90	V3	V3	V	V
9223	552	AAG	K	0	1	CAG	Q	1	0	1	0	1	0	1	1	5	0.0	-	66.60	V3	V3	V	V
9223	552	AAG	K	0	1	AAC	N	0	0	1	0	0	0	1	1	5	-2.0	-	*69.60	V4	V4	V	V
9667	700	TGG	W	0	2	TTC	F	0	0	0	0	0	0	1	1	0	0.0	-	66.60	V4	V4	V	N
9667	700	TGG	W	0	2	GCG	A	0	0	0	0	0	0	1	1	0	-8.0	-	18.50	V4	V4	V	N
9667	700	TGG	W	1	1	TAC	Y	0	0	0	0	0	0	1	1	0	1.0	-	69.00	V4	V4	V	N
9667	700	TGG	W	1	2	GAC	D	0	0	0	0	0	0	0	1	+0	-8.0	-	-	N6	N6	N	N
9667	700	TGG	W	1	1	GAG	E	0	0	0	0	0	0	1	1	+0	-6.0	-	49.70	V4	V4	V	N
9667	700	TGG	W	0	1	TGC	C	0	0	0	0	0	0	0	1	0	-5.0	-	-	N6	N6	N	N
9700	711	CAC	H	0	1	AAC	N	0	0	0	0	0	0	1	1	7	1.1	-	75.60	V4	V4	V	N
9700	711	CAC	H	1	0	CGC	R	0	0	0	0	0	0	0	1	+7	-2.0	-	-	N6	N6	N	N

Table 3: Snapshot of results generated by the GeneSV system for 32 hypothetical mutations (complete output is provided in the supplemental data – Suppl 1), experimental validation of those results, and comparison of those results to other similar systems. The mutations listed in blue font are those that were experimentally confirmed as not viable. The mutations listed in green font are those with a correct prediction of their

functionality. The mutations listed in red font are those with an incorrect prediction of their functionality. Viable prediction are marked as (V), not viable (N), and by (-) no results reported. For each base position the following information is reported:

Rnum: residue number within a reference protein
Cod2: codon at the base position in the reference
A: corresponding amino acid
TsTv: number of transitions transversions for Cod2 -> Cod1 mutation
Cod1: codon at the base position in the test (mutant) set
i: binary check if a given codon fulfills criteria Oi (i=1,..,8),
ch: conservation of charged amino acids (HKR DE) (0-9); '+' when test is charged
sCON: sequence conservation [-9,9]
gcSid4: sequence identity to genomic sequences from the expanded dataset (exact codon match)
aaSidH: sequence identity to protein sequences from homologous proteins only (amino acid match)
GSV1: viability predictions V1–V4 and N5-N7 calculated by GeneSV using dataset 2012.02.12
GSV2: viability predictions V1–V4 and N5-N7 calculated by GeneSV using dataset 2013.02.18
SIFT: viability predictions V or N by SIFT server
PRV: viability predictions V or N by PROVEAN server
(*) - codons with fewer than 3 position hits should be evaluated with caution (“viability” predictions may be not reliable).

Mutation in infectious clone	GeneSV Prediction	Virus Rescued	Prediction Validated?
(AAA)K-279-N(AAC) **(AGA)R-437-E(GAA)**	Viable	Not Viable	No
(AAA)K-279-R(AGA) **(AGA)R-437-E(GAA)**	Viable	Yes	Yes
(AAA)K-279-E(GAA) **(AGA)R-437-K(AAA)**	Viable	Yes	Yes
(AAA)K-279-E(GAA) **(AGA)R-437-Q(CAA)**	Viable	Yes	Yes
(AAA)K-279-E(GAA) **(AGA)R-437-E(GAA)**	Not Viable	Not Viable	Yes

Table 4: Results from the experimental validation of 5 sets of double mutations at the positions K279 and R437 for which constructed mutant viruses were rescued from plasmid transfected in Vero cells.

Method	**Datasets for homology searches**			**Conservation analysis**	
	Genomic sequences	Protein sequences	Protein structures	Sequence alignment	Structural features
SIFT	-	yes	-	PsiBlast	-
PROVEAN	-	yes	-	Blastp	-
PolyPhen-2	-	yes	yes	Blast MAFFT	yes* b,c
GeneSV	yes	yes	yes	PsiBlast Smith-Waterman	yes a,b,c,d,e

Table 5: Comparison of sequence mutation analysis tools based on types of datasets used for homology searches, and approaches they use for conservation analysis. List of structural features contributing to calculated conservation scores:

a. accuracy of constructed structural models
b. surface exposure (location of the mutation position: buried, exposed)
c. structural element (coil, helix, strand)
d. antigenic site (position within the epitope: yes, not)
e. structural conservation

(*) – structure-based features are estimated from the analysis of closest homologs with known 3D structure (no structural models are constructed).

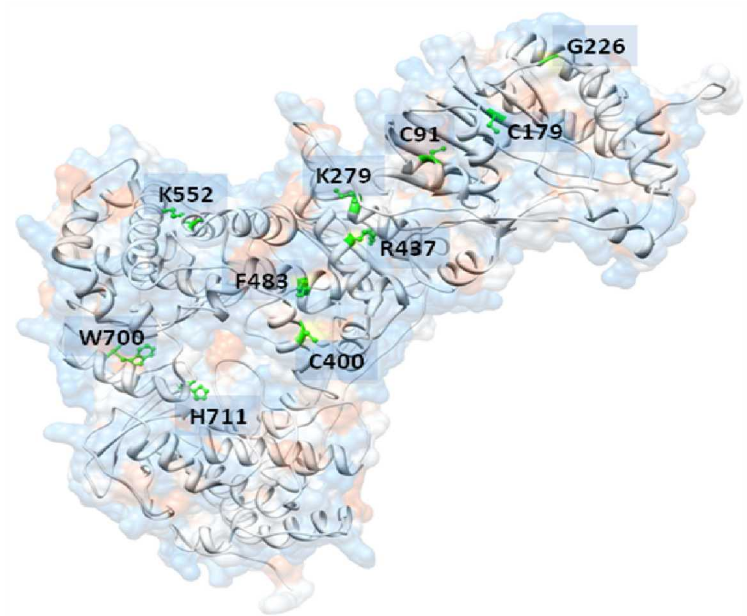

Figure 1. A structural model of NS5 polymerase (FLIC; gi:132271146; 900 aa). Residues highlighted in green are those positions that were selected for testing various predictions via generation of mutants.

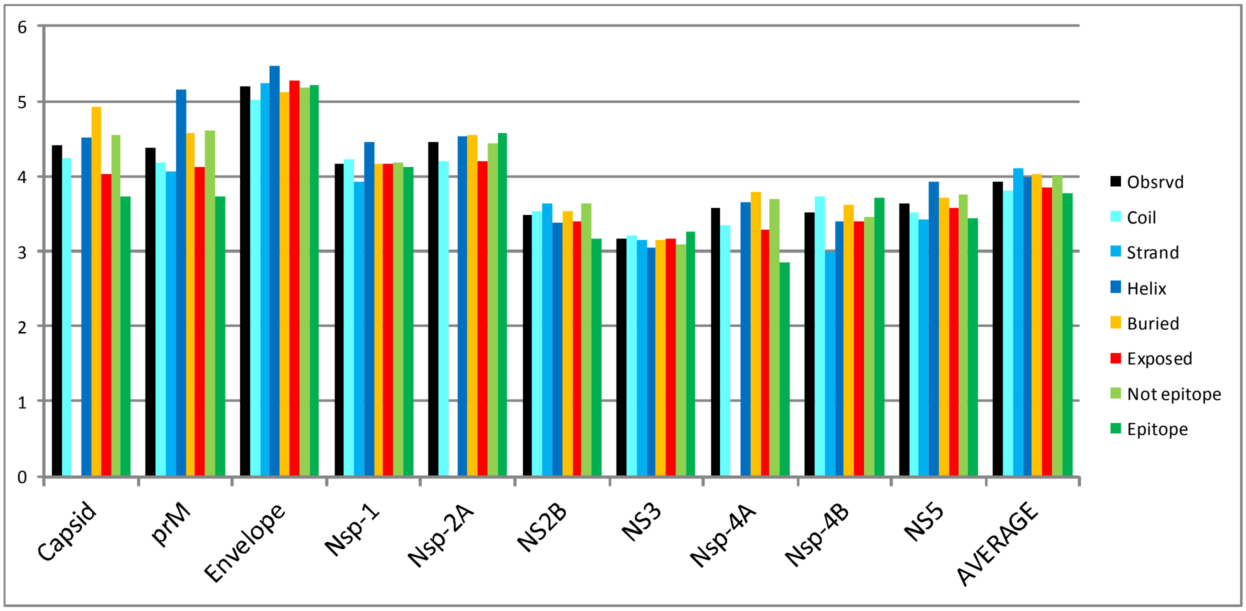

Figure 2. Column chart representation of the frequencies of mutations (codon variations) observed in the locations with different amino-acid and position characteristics within each protein coding region in DENV-2 dataset 2013.02.18. The bars colored in black and marked as "Obsrvd" show average frequencies observed within a given gene. Frequencies of mutations observed in the coil regions within a given gene are colored in cyan, strand – blue, helix – dark blue, buried – yellow, exposed – red, not a part of the antigenic epitopes – light green, and the frequencies of mutations within the epitope regions – green. In the AVERAGE set of bars are shown corresponding frequencies calculated as average from all combined protein coding regions.